In the directory for a tumor sample (trios will have two tumor directories) you will find:

- The VCF file
    - For pilot data this includes the verification results (VCF file with *_annotated_* in the name)
- The Full MAF file
    - Averages 4 million rows, includes all Germline, Somatic, and LOH variants
    - Not possible to open in Excel
- The "Somatic" MAF file
    - Includes only Somatic and LOH variant calls, generated by same CGI script the produces full MAF
    - Usually small enough to open in Excel
- The "Chromosome" MAFs
    - Generated by DCC from the full MAF to produce smaller files easier to parse
    - Can be opened in Excel
- The "verified_data_only" MAF data
    - Generated by DCC to give a summary of ANY data in MAF file that showed up as CGI lab verified in some way
    - Very small, can be opened in Excel

There is also a file named "results_from_parsed" in every tumor directory, this is a file used in DCC analysis (contains some parsed information from the MAF file) but will likely not be of interest to you, please just disregard this file.

In addition the DCC has produced some "summary" files across the whole disease, described below. The less interesting files are in blue (these files are more for a general idea of how the samples break down across the disease and are useful for QC), the more interesting files are shown in red. All files are tab separated values (.tsv) which can easily be opened in/converted to Excel.

- total_variant_summary_DISEASE.tsv – quick look at how whole MAF file breaks down into Somatic, LOH, and Germline variant calls
- variants_in_genes_summary_DISEASE.tsv – analysis of how the variants that fall in genes break down by variant type (missense, intron, etc.)
- variants_in_genes_somatic_and_loh_summary_DISEASE.tsv – same as above file (variants_in_genes_summary_DISEASE.tsv), restricted to somatic and LOH variants (Germline variants excluded)
- all_verified_variants_details_DISEASE.tsv – summary of all verified MAF data for whole disease, basically a concatenation of all "verified_data_only" files from each tumor in the disease (currently only available for pilot data). The last 8 columns of the MAF file contain lab verification information.
- variants_in_genes_somatic_and_loh_non_silent_details_DISEASE.tsv – **Please note that I feel that this file contains the most useful balance of details on interesting variants (Somatic and LOH, in genes, non-silent) across the entire disease and yet is still a manageable size file to look at in Excel. This is a good starting point for analysis that goes beyond just the bioinformatically and lab verified somatic variants (next two files).** This is a MAF file containing all Somatic and LOH variants from the disease that fall within genes and are "non-silent". This file matches the format of the MAF file exactly except for two things: first, an additional "Gene

Function" column has been added at end of the MAF columns, and unlike the original MAFs where multiple HUGO gene_syms can be in a single row (a single variant can technically fall within multiple genes) this file pulls them out and only has one gene possible per line, making analysis on a gene by gene level a little easier (i.e., you can more easily sort by gene_sym in this file). If a single Somatic or LOH variant had multiple gene_syms that were non-silent the data would all show up, just not in the same row as they would in the original MAF."In gene" defined as any variant contained within a gene by CGI (i.e., the HUGO Symbol is filled out in the MAF for this variant, includes intron/exon/UTR and 7.5 kbp upstream of gene). "Non-silent" defined as any of the following Variant Classifications (see MAF file README page 2 for definitions of Variant Classifications):

  - NONSTOP
  - MISSTART
  - NONSENSE
  - MISSENSE
  - SPAN5
  - SPAN3
  - SPAN
  - INSERT
  - DELETE
  - INSERT+
  - DELETE+
  - DISRUPT
  - FRAMESHIFT
  - ACCEPTOR
  - DONOR

  Also available as an Excel (.xlsx) file.
- sqhigh_and_verified_variants_in_genes_somatic_non_silent_details_DISEASE.tsv – A subset of the above file (variants_in_genes_somatic_and_loh_non_silent_details_DISEASE.tsv) includes ***only*** variants that are somatic, fall within a gene, non-silent, and either marked "SQHIGH" in the Somatic_quality column (i.e., are "bioinformatically" verified) OR were called "Somatic" during lab verification. Also available as an Excel (.xlsx) file.
- sqhigh_and_verified_variants_in_genes_somatic_non_silent_summary_with_filters_DISEASE.tsv – Exact same data as last file, except summarized across the disease to show which genes got multiple variants and processed through some filters to try and determine the highest quality hits. **Currently the two filters eliminate anything in the previous file with a somatic _rank of less than 0.1 AND an FET_score of less than 13.** Additional filters can be added at the Project Team's request. This file is a very quick way to find the genes that were hit multiple times by bioinformatically or lab verified somatic non-silent variants. Also available as an Excel (.xlsx) file.

Inter-disease comparisons:

sqhigh_and_verified_variants_in_genes_somatic_non_silent_summary_all_diseases_with_filters_For_DISEASE.xlsx - A summary of the findings across all 5 TARGET diseases. The information from the first 3 columns of the above file (sqhigh_and_verified_variants_in_genes_somatic_non_silent_summary_with_filters_DISEASE.tsv) will match up to your disease's data in this file. Disease information from other 4 TARGET diseases have been anonymized.